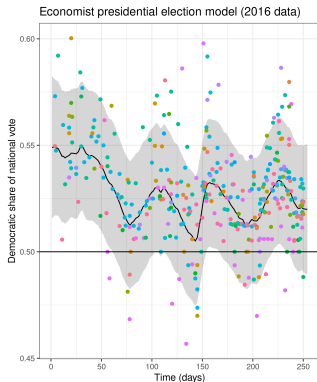


Approximate data deletion and replication with the Bayesian influence function

Ryan Giordano (rgiordano@berkeley.edu, UC Berkeley), Tamara Broderick (MIT)
Stanford Statistics Seminar May 2024

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

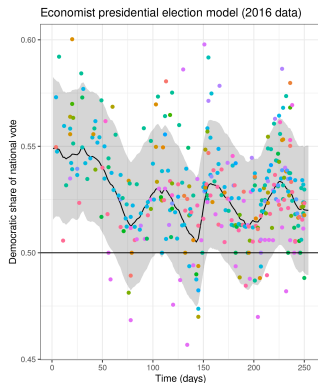
Model:

- $X = x_1, \dots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\mathbb{E}_{p(\theta|X)} [f(\theta)]$.

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \dots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

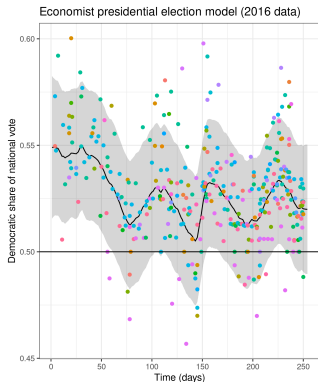
We want to know $\mathbb{E}_{p(\theta|X)} [f(\theta)]$.

The people who responded to the polls were randomly selected.

If we had selected a different random sample, how much would our estimate have changed?

Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

Economist 2016 Election Model [Gelman and Heidemanns, 2020]



A time series model to predict the 2016 US presidential election outcome from polling data.

Model:

- $X = x_1, \dots, x_N =$ Polling data ($N = 361$).
- $\theta =$ Lots of random effects (day, pollster, etc.)
- $f(\theta) =$ Democratic % of vote on election day

Typically, we compute Markov chain Monte Carlo (MCMC) draws from the posterior $p(\theta|X)$.

We want to know $\mathbb{E}_{p(\theta|X)} [f(\theta)]$.

The people who responded to the polls were randomly selected.

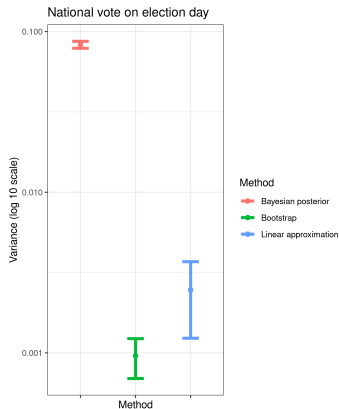
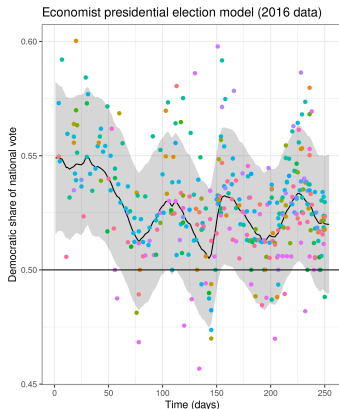
If we had selected a different random sample, how much would our estimate have changed?

Idea: Re-fit with bootstrap samples of data [Huggins and Miller, 2023]

Problem: Each MCMC run takes about 10 hours (Stan, six cores).

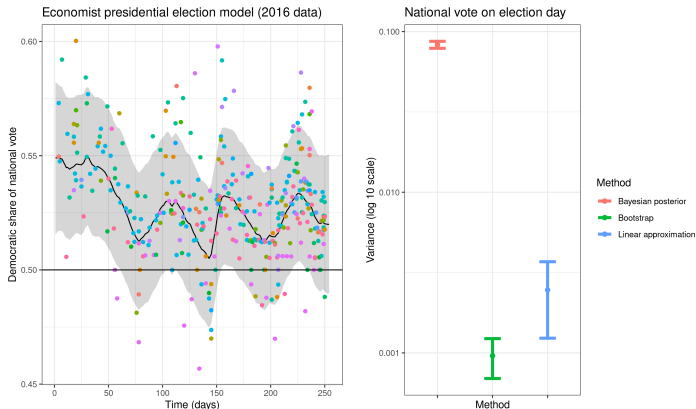
Results

Proposal: Use full-data posterior draws to form a linear approximation to *data reweightings*.



Results

Proposal: Use full-data posterior draws to form a linear approximation to *data reweightings*.



Compute time for 100 bootstraps: 51 days

Compute time for the linear approximation: Seconds
(But note the approximation has some error)

- Data reweighting
 - Write the change in the posterior expectation as **linear component** + **error**
 - The **linear component** can be computed from a single run of MCMC

- Data reweighting
 - Write the change in the posterior expectation as **linear component** + **error**
 - The **linear component** can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
 - As $N \rightarrow \infty$, the linear component provides an arbitrarily good approximation
 - Consistent variance estimates via a uniform Bernstein–von Mises theorem

- Data reweighting
 - Write the change in the posterior expectation as **linear component** + **error**
 - The **linear component** can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
 - As $N \rightarrow \infty$, the linear component provides an arbitrarily good approximation
 - Consistent variance estimates via a uniform Bernstein–von Mises theorem
- High-dimensional problems
 - The linear component is the same order as the error
 - Even for parameters which concentrate, even as $N \rightarrow \infty$
 - Study the variance estimates via a Bayesian von–Mises expansion

- Data reweighting
 - Write the change in the posterior expectation as **linear component** + **error**
 - The **linear component** can be computed from a single run of MCMC
- Finite-dimensional problems with posteriors which concentrate asymptotically
 - As $N \rightarrow \infty$, the linear component provides an arbitrarily good approximation
 - Consistent variance estimates via a uniform Bernstein–von Mises theorem
- High-dimensional problems
 - The linear component is the same order as the error
 - Even for parameters which concentrate, even as $N \rightarrow \infty$
 - Study the variance estimates via a Bayesian von–Mises expansion
- Some implications and future work

Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta) \qquad \log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

Original weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

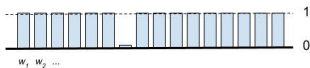
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

Original weights:



Leave-one-out weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

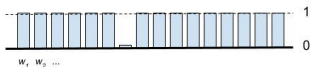
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

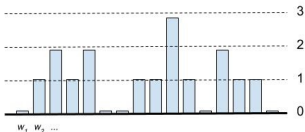
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

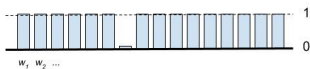
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

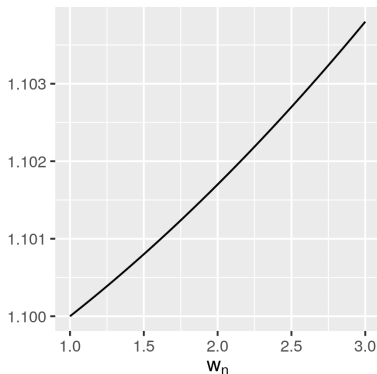
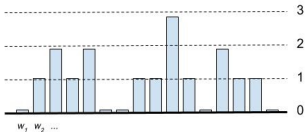
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

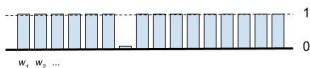
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

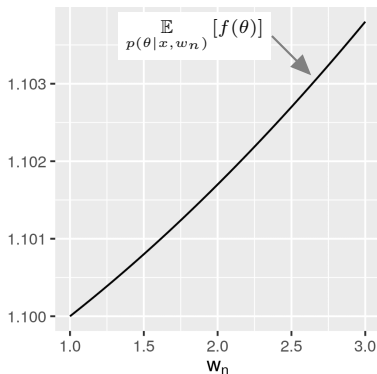
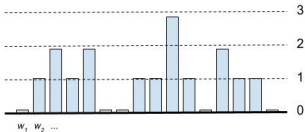
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

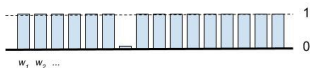
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

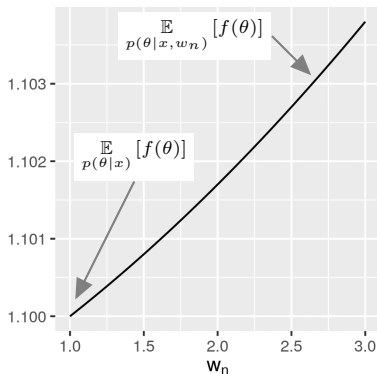
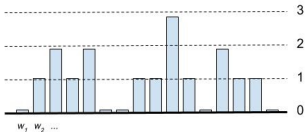
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

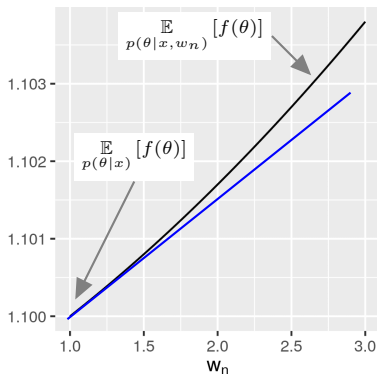
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

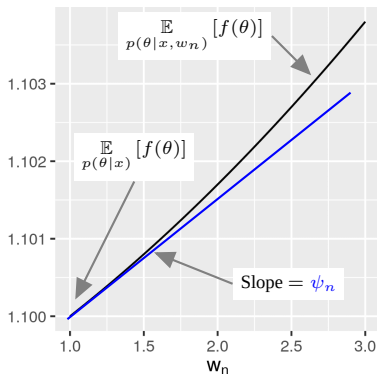
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

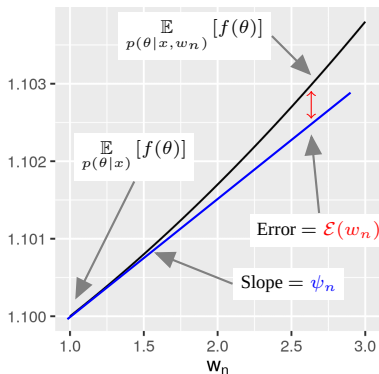
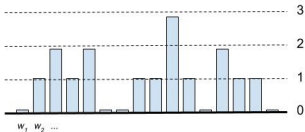
Original weights:



Leave-one-out weights:



Bootstrap weights:



Data re-weighting.

Augment the problem with *data weights* w_1, \dots, w_N . We can write $\mathbb{E}_{p(\theta|X,w)}[f(\theta)]$.

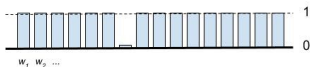
$$\ell_n(\theta) := \log p(x_n|\theta)$$

$$\log p(X|\theta, w) = \sum_{n=1}^N w_n \ell_n(\theta)$$

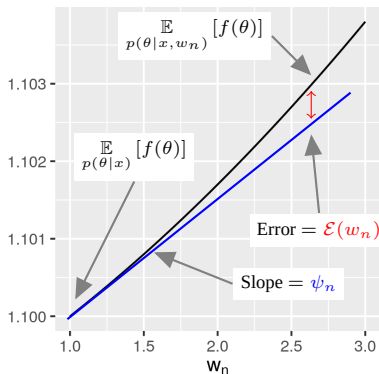
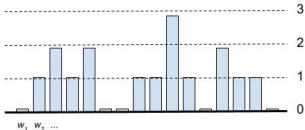
Original weights:



Leave-one-out weights:



Bootstrap weights:



The re-scaled slope $N\psi_n$ is known as the “influence function” at data point x_n .

$$\mathbb{E}_{p(\theta|X,w)}[f(\theta)] - \mathbb{E}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^N \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

How can we use the approximation?

Assume the **slope** is computable and **error** is small.

$$\mathbb{E}_{p(\theta|X,w)}[f(\theta)] - \mathbb{E}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^N \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

How can we use the approximation?

Assume the **slope** is computable and **error** is small.

$$\mathbb{E}_{p(\theta|X,w)}[f(\theta)] - \mathbb{E}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^N \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Bootstrap. Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

$$\text{Bootstrap variance} = \text{Var}_{p(w)} \left(\mathbb{E}_{p(\theta|X,w)}[f(\theta)] \right)$$

How can we use the approximation?

Assume the **slope** is computable and **error** is small.

$$\mathbb{E}_{p(\theta|X,w)}[f(\theta)] - \mathbb{E}_{p(\theta|X)}[f(\theta)] = \sum_{n=1}^N \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Bootstrap. Draw bootstrap weights $w \sim p(w) = \text{Multinomial}(N, N^{-1})$.

$$\begin{aligned} \text{Bootstrap variance} &= \text{Var}_{p(w)} \left(\mathbb{E}_{p(\theta|X,w)}[f(\theta)] \right) \\ &= \text{Var}_{p(w)} \left(\sum_{n=1}^N \psi_n(w_n - 1) + \mathcal{E}(w_n) \right) \\ &= \frac{1}{N^2} \sum_{n=1}^N \left(\psi_n - \bar{\psi} \right)^2 + \text{Term involving } \mathcal{E}(w_n) \text{ for } n = 1, \dots, N \\ &\approx \frac{1}{N} \underbrace{\left(\frac{1}{N} \sum_{n=1}^N \left(\psi_n - \bar{\psi} \right)^2 \right)}_{\text{"Infinitesimal jackknife variance estimate"}} \end{aligned}$$

Expressions for the slope and error

How to compute the slopes ψ_n ? How large is the error $\mathcal{E}(w)$?

For simplicity, let us consider a single weight for the moment.

$$\mathbb{E}_{p(\theta|X, w_n)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Expressions for the slope and error

How to compute the slopes ψ_n ? How large is the error $\mathcal{E}(w)$?

For simplicity, let us consider a single weight for the moment.

$$\mathbb{E}_{p(\theta|X, w_n)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Let an overbar denote “posterior–mean zero.” For example, $\bar{f}(\theta) := f(\theta) - \mathbb{E}_{p(\theta|X)} [f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\psi_n = \underbrace{\mathbb{E}_{p(\theta|X)} [\bar{f}(\theta) \bar{\ell}_n(\theta)]}_{\text{Estimatable with MCMC!}} \quad \mathcal{E}(w_n) = \frac{1}{2} \underbrace{\mathbb{E}_{p(\theta|X, \tilde{w}_n)} [\bar{f}(\theta) \bar{\ell}_n(\theta) \bar{\ell}_n(\theta)]}_{\text{Cannot compute directly (don't know } \tilde{w})} (w_n - 1)^2$$

Expressions for the slope and error

How to compute the slopes ψ_n ? How large is the error $\mathcal{E}(w)$?

For simplicity, let us consider a single weight for the moment.

$$\mathbb{E}_{p(\theta|X, w_n)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \psi_n(w_n - 1) + \mathcal{E}(w_n)$$

Let an overbar denote “posterior–mean zero.” For example, $\bar{f}(\theta) := f(\theta) - \mathbb{E}_{p(\theta|X)} [f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\begin{aligned} \psi_n &= \underbrace{\mathbb{E}_{p(\theta|X)} [\bar{f}(\theta) \bar{\ell}_n(\theta)]}_{\text{Estimatable with MCMC!}} & \mathcal{E}(w_n) &= \frac{1}{2} \underbrace{\mathbb{E}_{p(\theta|X, \tilde{w}_n)} [\bar{f}(\theta) \bar{\ell}_n(\theta) \bar{\ell}_n(\theta)]}_{\text{Cannot compute directly (don't know } \tilde{w})} (w_n - 1)^2 \\ &= O_p(N^{-1}) \text{ under posterior concentration} & &= O_p(N^{-2}) \text{ under posterior concentration} \end{aligned}$$

Expressions for the slope and error

How to compute the slopes ψ_n ? How large is the error $\mathcal{E}(w)$?

For simplicity, let us consider a single weight for the moment.

$$\mathbb{E}_{p(\theta|X, w_n)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] = \psi_n (w_n - 1) + \mathcal{E}(w_n)$$

Let an overbar denote “posterior–mean zero.” For example, $\bar{f}(\theta) := f(\theta) - \mathbb{E}_{p(\theta|X)} [f(\theta)]$.

By dominated convergence and the mean value theorem, for some $\tilde{w}_n \in [0, w_n]$:

$$\begin{aligned} \psi_n &= \underbrace{\mathbb{E}_{p(\theta|X)} [\bar{f}(\theta) \bar{\ell}_n(\theta)]}_{\text{Estimatable with MCMC!}} & \mathcal{E}(w_n) &= \frac{1}{2} \underbrace{\mathbb{E}_{p(\theta|X, \tilde{w}_n)} [\bar{f}(\theta) \bar{\ell}_n(\theta) \bar{\ell}_n(\theta)]}_{\text{Cannot compute directly (don't know } \tilde{w})} (w_n - 1)^2 \\ &= O_p(N^{-1}) \text{ under posterior concentration} & &= O_p(N^{-2}) \text{ under posterior concentration} \end{aligned}$$

Theorem 1 [Giordano and Broderick, 2023] (paraphrase):

If the posterior $p(\theta|X)$ “concentrates” (e.g. as in the Bernstein–von Mises theorem),^a then

$$w_n \mapsto N \left(\mathbb{E}_{p(\theta|X, w_n)} [f(\theta)] - \mathbb{E}_{p(\theta|X)} [f(\theta)] \right)$$

becomes linear as $N \rightarrow \infty$, with slope $\lim_{N \rightarrow \infty} \psi_n$.

^aExisting results are sufficient for a *particular weight* [Kass et al., 1990]. Giordano and Broderick [2023] proves that the result holds when averaged over all weights, as needed for variance estimation.

Negative binomial experiment

Example: Negative binomial models with an unknown parameter γ .

For $n = 1, \dots, N$ let $x_n | \gamma \stackrel{iid}{\sim} \text{NegativeBinomial}(\alpha, \gamma)$ for fixed α .

$$\text{Write } \log p(X | \lambda, \gamma, w) = \sum_{n=1}^N w_n \ell_n(\gamma).$$

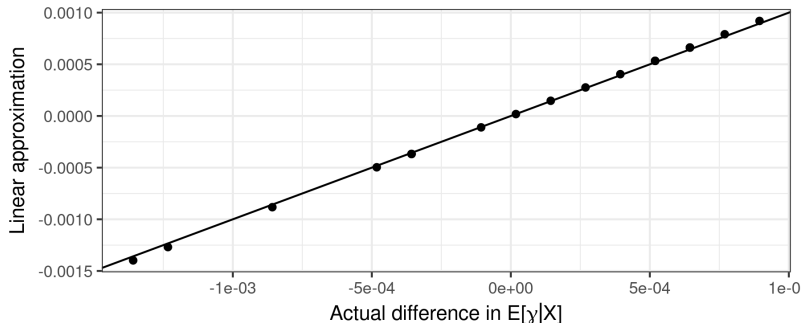
Negative binomial experiment

Example: Negative binomial models with an unknown parameter γ .

For $n = 1, \dots, N$ let $x_n | \gamma \stackrel{iid}{\sim} \text{NegativeBinomial}(\alpha, \gamma)$ for fixed α .

$$\text{Write } \log p(X | \lambda, \gamma, w) = \sum_{n=1}^N w_n \ell_n(\gamma).$$

Negative Binomial model
leaving out single datapoints with $N = 800$



Variance consistency theorem

How do the results for a single weight translate into variance estimates?

$$\text{Var}_{p(w)} \left(\mathbb{E}_{p(\theta|X,w)} [f(\theta)] \right) = \frac{1}{N^2} \sum_{n=1}^N \left(\psi_n - \bar{\psi} \right)^2 + \text{Term involving } \mathcal{E}(w_n) \text{ for } n = 1, \dots, N$$

Variance consistency theorem

How do the results for a single weight translate into variance estimates?

$$\text{Var}_{p(w)} \left(\mathbb{E}_{p(\theta|X,w)} [f(\theta)] \right) = \frac{1}{N^2} \sum_{n=1}^N \left(\psi_n - \bar{\psi} \right)^2 + \text{Term involving } \mathcal{E}(w_n) \text{ for } n = 1, \dots, N$$

- Assume: A well-behaved MAP *maximum a posteriori* estimator $\hat{\theta}$ exists.
 - The dimension of θ is fixed as $N \rightarrow \infty$.
 - The expected log likelihood has a unique maximum at θ_∞
 - The observed log likelihood satisfies $\hat{\theta} \rightarrow \theta_\infty$
 - The expected log likelihood Hessian \mathcal{I} is negative definite at θ_∞
- Assume: We can apply standard asymptotics.
 - The log prior and log likelihood are four times continuously differentiable
 - The prior is proper, and a technical set of squared expectations are finite
 - The log likelihood derivatives are dominated by a square-integrable envelope function in a neighborhood of θ_∞ .

Variance consistency theorem

How do the results for a single weight translate into variance estimates?

$$\text{Var}_{p(w)} \left(\mathbb{E}_{p(\theta|X, w)} [f(\theta)] \right) = \frac{1}{N^2} \sum_{n=1}^N \left(\psi_n - \bar{\psi} \right)^2 + \text{Term involving } \mathcal{E}(w_n) \text{ for } n = 1, \dots, N$$

- Assume: A well-behaved MAP *maximum a posteriori* estimator $\hat{\theta}$ exists.
 - The dimension of θ is fixed as $N \rightarrow \infty$.
 - The expected log likelihood has a unique maximum at θ_∞
 - The observed log likelihood satisfies $\hat{\theta} \rightarrow \theta_\infty$
 - The expected log likelihood Hessian \mathcal{I} is negative definite at θ_∞
- Assume: We can apply standard asymptotics.
 - The log prior and log likelihood are four times continuously differentiable
 - The prior is proper, and a technical set of squared expectations are finite
 - The log likelihood derivatives are dominated by a square-integrable envelope function in a neighborhood of θ_∞ .

Theorem 2 [Giordano and Broderick, 2023]: Under the above assumptions,

$$\sqrt{N} \left(\mathbb{E}_{p(\theta|X)} [g(\theta)] - g(\theta_\infty) \right) \xrightarrow[N \rightarrow \infty]{dist} \mathcal{N}(0, V^g) \quad [\text{Kleijn and Van der Vaart, 2012}]$$

$$\text{and } V^{\text{IJ}} := \frac{1}{N} \sum_{n=1}^N \left(\psi_n - \bar{\psi} \right)^2 \xrightarrow[N \rightarrow \infty]{prob} V^g. \quad (\text{our contribution})$$

Data Analysis Using Regression and Multilevel/Hierarchical Models.

We ran `rstanarm` on 56 different models on 13 different datasets from Gelman and Hill [2006], including Gaussian and logistic regression, fixed and mixed-effects models.

Across all models, we estimate 799 distinct covariances (regression coefficients and log scale parameters).

Using the bootstrap as ground truth, compute the relative errors:

$$\frac{V_{\text{Bayes}} - V_{\text{Boot}}}{|V_{\text{Boot}}|} \quad \text{and} \quad \frac{V_{\text{IJ}} - V_{\text{Boot}}}{|V_{\text{Boot}}|}.$$

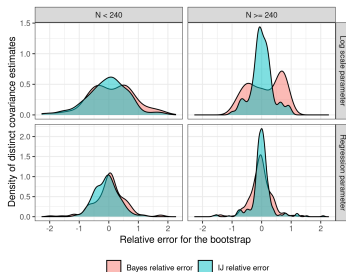


Figure 1: The distribution of the relative errors. Log scale parameters include all variances or covariances that involve at least one log scale parameters.

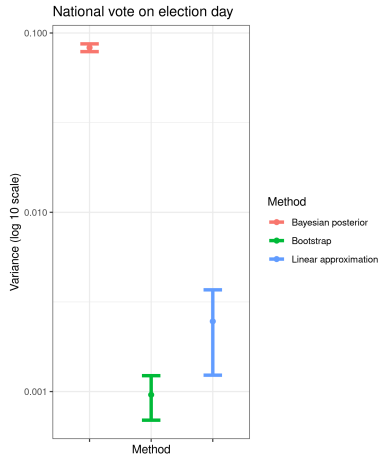
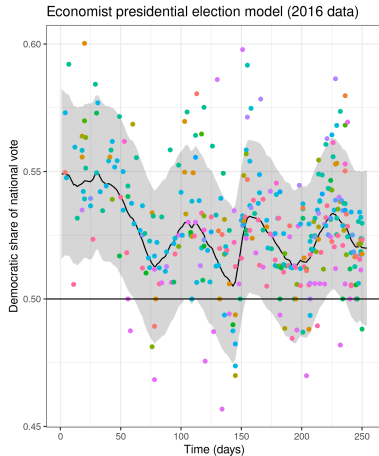
Total compute time for all models:

Initial fit: 1.6 hours

Bootstrap: 381.5 hours

How to connect to the election data?

Problem: MCMC is only interesting when the posterior doesn't concentrate.



Example: Exponential families with random effects (REs) λ and fixed effects γ .

If the observations per random effect remains bounded as $N \rightarrow \infty$, then

- Parameter λ (“local”) grows in dimension with N .
- Parameter γ (“global”) is finite-dimensional.
- Marginally $p(\lambda|X)$ does not concentrate.
- Marginally, $p(\gamma|X)$ concentrates.

High dimensional problems

Example: Exponential families with random effects (REs) λ and fixed effects γ .

If the observations per random effect remains bounded as $N \rightarrow \infty$, then

- Parameter λ (“local”) grows in dimension with N .
- Parameter γ (“global”) is finite-dimensional.
- Marginally $p(\lambda|X)$ does not concentrate.
- Marginally, $p(\gamma|X)$ concentrates.

In general, we cannot hope for an asymptotic analysis of $\mathbb{E}_{p(\lambda, \gamma|X)} [f(\lambda)]$.

Can we save the approximation when *some* parameters concentrate?

Does the residual vanish asymptotically for $w_n \mapsto \mathbb{E}_{p(\gamma|X, w_n)} [f(\gamma)]$?

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$\begin{aligned} \mathbb{E}_{p(\gamma, \lambda|X, w_n)}[\gamma] - \mathbb{E}_{p(\gamma, \lambda|X)}[\gamma] = \\ \psi_n(w_n - 1) + \mathcal{E}(w_n) \end{aligned}$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$\begin{aligned} & \mathbb{E}_{p(\gamma, \lambda|X, w_n)} [\gamma] - \mathbb{E}_{p(\gamma, \lambda|X)} [\gamma] = \\ & \quad \psi_n(w_n - 1) + \mathcal{E}(w_n) \\ = & \mathbb{E}_{p(\gamma, \lambda|X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)] (w_n - 1) + \frac{1}{2} \mathbb{E}_{p(\gamma, \lambda|X, \tilde{w}_n)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)^2] (w_n - 1)^2 \end{aligned}$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$\begin{aligned}
 & \mathbb{E}_{p(\gamma, \lambda|X, w_n)} [\gamma] - \mathbb{E}_{p(\gamma, \lambda|X)} [\gamma] = \\
 & \quad \psi_n(w_n - 1) + \mathcal{E}(w_n) \\
 = & \mathbb{E}_{p(\gamma, \lambda|X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)] (w_n - 1) + \frac{1}{2} \mathbb{E}_{p(\gamma, \lambda|X, \tilde{w}_n)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)^2] (w_n - 1)^2 \\
 = & \mathbb{E}_{p(\gamma|X)} \left[\bar{\gamma} \underbrace{\mathbb{E}_{p(\lambda|\gamma, X)} [\bar{\ell}_n(\gamma, \lambda)]}_{F_1(\gamma)} \right] (w_n - 1) + \frac{1}{2} \mathbb{E}_{p(\gamma|X, \tilde{w}_n)} \left[\bar{\gamma} \underbrace{\mathbb{E}_{p(\lambda|X, \gamma, \tilde{w}_n)} [\bar{\ell}_n(\gamma, \lambda)^2]}_{F_2(\gamma)} \right] (w_n - 1)
 \end{aligned}$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$\begin{aligned}
 & p(\gamma, \lambda|X, w_n) \mathbb{E}[\gamma] - p(\gamma, \lambda|X) \mathbb{E}[\gamma] = \\
 & \quad \psi_n(w_n - 1) + \mathcal{E}(w_n) \\
 & = \mathbb{E}_{p(\gamma, \lambda|X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)] (w_n - 1) + \frac{1}{2} \mathbb{E}_{p(\gamma, \lambda|X, \tilde{w}_n)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)^2] (w_n - 1)^2 \\
 & = \mathbb{E}_{p(\gamma|X)} \left[\bar{\gamma} \underbrace{\mathbb{E}_{p(\lambda|\gamma, X)} [\bar{\ell}_n(\gamma, \lambda)]}_{F_1(\gamma)} \right] (w_n - 1) + \frac{1}{2} \mathbb{E}_{p(\gamma|X, \tilde{w}_n)} \left[\bar{\gamma} \underbrace{\mathbb{E}_{p(\lambda|X, \gamma, \tilde{w}_n)} [\bar{\ell}_n(\gamma, \lambda)^2]}_{F_2(\gamma)} \right] (w_n - 1)^2 \\
 & = \underbrace{\mathbb{E}_{p(\gamma|X)} [\bar{\gamma} F_1(\gamma)]}_{O_p(N^{-1})} (w_n - 1) + \underbrace{\frac{1}{2} \mathbb{E}_{p(\gamma|X, \tilde{w}_n)} [\bar{\gamma} F_2(\gamma)]}_{O_p(N^{-1})} (w_n - 1)^2 \\
 & \quad \text{(by } p(\gamma|X) \text{ concentration)} \quad \text{(by } p(\gamma|X) \text{ concentration)} \\
 & \Rightarrow
 \end{aligned}$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$\begin{aligned}
 & p(\gamma, \lambda|X, w_n) \mathbb{E}[\gamma] - p(\gamma, \lambda|X) \mathbb{E}[\gamma] = \\
 & \quad \psi_n(w_n - 1) + \mathcal{E}(w_n) \\
 = & \quad \mathbb{E}_{p(\gamma, \lambda|X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)] (w_n - 1) + \frac{1}{2} \mathbb{E}_{p(\gamma, \lambda|X, \tilde{w}_n)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)^2] (w_n - 1)^2 \\
 = & \quad \mathbb{E}_{p(\gamma|X)} \left[\bar{\gamma} \underbrace{\mathbb{E}_{p(\lambda|\gamma, X)} [\bar{\ell}_n(\gamma, \lambda)]}_{F_1(\gamma)} \right] (w_n - 1) + \frac{1}{2} \mathbb{E}_{p(\gamma|X, \tilde{w}_n)} \left[\bar{\gamma} \underbrace{\mathbb{E}_{p(\lambda|X, \gamma, \tilde{w}_n)} [\bar{\ell}_n(\gamma, \lambda)^2]}_{F_2(\gamma)} \right] (w_n - 1) \\
 = & \quad \underbrace{\mathbb{E}_{p(\gamma|X)} [\bar{\gamma} F_1(\gamma)]}_{O_p(N^{-1})} (w_n - 1) + \frac{1}{2} \underbrace{\mathbb{E}_{p(\gamma|X, \tilde{w}_n)} [\bar{\gamma} F_2(\gamma)]}_{O_p(N^{-1})} (w_n - 1)^2 \\
 & \quad \text{(by } p(\gamma|X) \text{ concentration)} \quad \text{(by } p(\gamma|X) \text{ concentration)} \\
 \Rightarrow & \quad \psi_n = O_p(N^{-1}) \quad \mathcal{E}(w_n) = O_p(N^{-1})
 \end{aligned}$$

High dimensional problems

We assume that $p(\gamma|X)$ concentrates but $p(\lambda|X)$ does not. By our series expansion:

$$\begin{aligned}
 & p(\gamma, \lambda|X, w_n) [\gamma] - p(\gamma, \lambda|X) [\gamma] = \\
 & \quad \psi_n(w_n - 1) + \mathcal{E}(w_n) \\
 = & \quad \mathbb{E}_{p(\gamma, \lambda|X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)] (w_n - 1) + \frac{1}{2} \mathbb{E}_{p(\gamma, \lambda|X, \tilde{w}_n)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)^2] (w_n - 1)^2 \\
 = & \quad \mathbb{E}_{p(\gamma|X)} \left[\bar{\gamma} \underbrace{\mathbb{E}_{p(\lambda|\gamma, X)} [\bar{\ell}_n(\gamma, \lambda)]}_{F_1(\gamma)} \right] (w_n - 1) + \frac{1}{2} \mathbb{E}_{p(\gamma|X, \tilde{w}_n)} \left[\bar{\gamma} \underbrace{\mathbb{E}_{p(\lambda|X, \gamma, \tilde{w}_n)} [\bar{\ell}_n(\gamma, \lambda)^2]}_{F_2(\gamma)} \right] (w_n - 1)^2 \\
 = & \quad \underbrace{\mathbb{E}_{p(\gamma|X)} [\bar{\gamma} F_1(\gamma)]}_{O_p(N^{-1})} (w_n - 1) + \frac{1}{2} \underbrace{\mathbb{E}_{p(\gamma|X, \tilde{w}_n)} [\bar{\gamma} F_2(\gamma)]}_{O_p(N^{-1})} (w_n - 1)^2 \\
 & \quad \text{(by } p(\gamma|X) \text{ concentration)} \quad \text{(by } p(\gamma|X) \text{ concentration)} \\
 \Rightarrow & \quad \psi_n = O_p(N^{-1}) \quad \mathcal{E}(w_n) = O_p(N^{-1})
 \end{aligned}$$

Corollary [Giordano and Broderick, 2023]:

In general, $w_n \mapsto N \left(\mathbb{E}_{p(\gamma|X, w_n)} [\gamma] - \mathbb{E}_{p(\gamma|X)} [\gamma] \right)$ remains non-linear as $N \rightarrow \infty$.

Example: Poisson regression with Gamma-distributed random effects

For $g = 1, \dots, G$, $\lambda_g \stackrel{iid}{\sim} \text{Gamma}(\alpha, \beta)$ for fixed α, β

For $n = 1, \dots, N$, $g_n \stackrel{iid}{\sim} \text{Categorical}(1, \dots, G)$, $y_n | \lambda_n, \gamma, g_n \stackrel{iid}{\sim} \text{Poisson}(\gamma \lambda_{g_n})$.

$x_n = (y_n, g_n)$ are IID given λ, γ . Write $\log p(X | \lambda, \gamma, w) = \sum_{n=1}^N w_n \ell_n(\lambda, \gamma)$.

Experiments

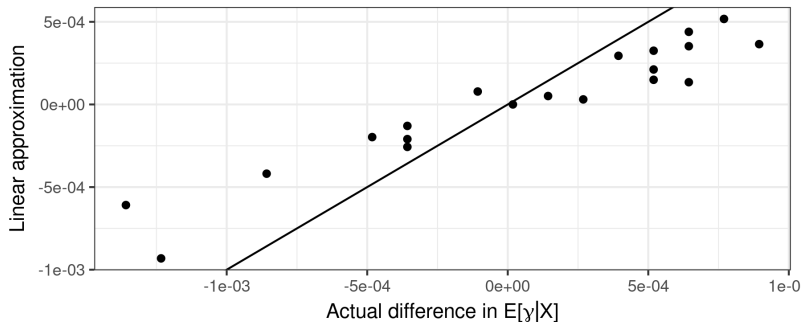
Example: Poisson regression with Gamma-distributed random effects

For $g = 1, \dots, G$, $\lambda_g \stackrel{iid}{\sim} \text{Gamma}(\alpha, \beta)$ for fixed α, β

For $n = 1, \dots, N$, $g_n \stackrel{iid}{\sim} \text{Categorical}(1, \dots, G)$, $y_n | \lambda_{g_n}, \gamma, g_n \stackrel{iid}{\sim} \text{Poisson}(\gamma \lambda_{g_n})$.

$x_n = (y_n, g_n)$ are IID given λ, γ . Write $\log p(X | \lambda, \gamma, w) = \sum_{n=1}^N w_n \ell_n(\lambda, \gamma)$.

Poisson random effect model
leaving out single datapoints with $N = 800$



How can we apply the single-weight result to variance computations?

Bayesian von-Mises Expansion

How can we apply the single-weight result to variance computations?

Define the “generalized posterior” functional

$$T(\mathbb{G}, N) := \frac{\int g(\theta) \exp \left(N \int \ell(x_0|\theta) \mathbb{G}(dx_0) \right) \pi(\theta) d\theta}{\int \exp \left(N \int \ell(x_0|\theta) \mathbb{G}(dx_0) \right) \pi(\theta) d\theta}.$$

Let \mathbb{F}_N denote the empirical distribution. Then

$$\mathbb{E}_{p(\theta|X)}[g(\theta)] = \frac{\int g(\theta) \exp \left(N \frac{1}{N} \sum_{n=1}^N \ell(x_n|\theta) \right) \pi(\theta) d\theta}{\int \exp \left(N \frac{1}{N} \sum_{n=1}^N \ell(x_n|\theta) \right) \pi(\theta) d\theta} = T(\mathbb{F}_N, N).$$

Bayesian von-Mises Expansion

How can we apply the single-weight result to variance computations?

Define the “generalized posterior” functional

$$T(\mathbb{G}, N) := \frac{\int g(\theta) \exp \left(N \int \ell(x_0|\theta) \mathbb{G}(dx_0) \right) \pi(\theta) d\theta}{\int \exp \left(N \int \ell(x_0|\theta) \mathbb{G}(dx_0) \right) \pi(\theta) d\theta}.$$

Let \mathbb{F}_N denote the empirical distribution. Then

$$\mathbb{E}_{p(\theta|X)}[g(\theta)] = \frac{\int g(\theta) \exp \left(N \frac{1}{N} \sum_{n=1}^N \ell(x_n|\theta) \right) \pi(\theta) d\theta}{\int \exp \left(N \frac{1}{N} \sum_{n=1}^N \ell(x_n|\theta) \right) \pi(\theta) d\theta} = T(\mathbb{F}_N, N).$$

Let \mathbb{F} denote the true distribution of x_n , and let $\mathbb{F}_N^t = t\mathbb{F}_N + (1-t)\mathbb{F}$.

We can study the *von Mises expansion*:

$$\begin{aligned} \sqrt{N} \left(\mathbb{E}_{p(\theta|X)}[g(\theta)] - T(\mathbb{F}, N) \right) &= \sqrt{N} \left. \frac{\partial T(\mathbb{F}_N^t, N)}{\partial t} \right|_{t=0} (\mathbb{F}_N - \mathbb{F}) + \mathcal{E}(\tilde{t}) \\ &= \underbrace{\sqrt{N} \sum_{n=1}^N (\psi_n - \bar{\psi})}_{\text{Infinitesimal jackknife estimator}} + o_p(1) + \mathcal{E}(\tilde{t}). \end{aligned}$$

Inconsistency is suggested if $\mathcal{E}(\tilde{t})$ fails to vanish.

Theorem 3 [Giordano and Broderick, 2023] (sketch):

(Consistency of the von-Mises expansion in finite dimensions)

Under slightly stronger conditions our original finite-dimensional posterior consistency result,

$$\sup_{\tilde{t} \in [0,1]} |\mathcal{E}(\tilde{t})| \rightarrow 0 \quad \text{in the Bayesian von-Mises expansion.}$$

Bayesian von-Mises Expansion Results

Theorem 3 [Giordano and Broderick, 2023] (sketch):

(Consistency of the von-Mises expansion in finite dimensions)

Under slightly stronger conditions our original finite-dimensional posterior consistency result,

$$\sup_{\tilde{t} \in [0,1]} |\mathcal{E}(\tilde{t})| \rightarrow 0 \quad \text{in the Bayesian von-Mises expansion.}$$

Theorem 4 [Giordano and Broderick, 2023] (sketch, not yet on arxiv):

(Inconsistency of the von-Mises expansion in infinite dimensions)

Assume that x_n comes with a random group assignment $g_n \in 1, \dots, G$. Conditional on g, x_n is modeled as a finite-dimensional exponential family given λ, γ :

$$\log p(x_n | g_n = g, \gamma, \lambda) = \tau(x_n)^\top \eta_g(\gamma, \lambda) + \text{Constant.}$$

Define the average product of second moments:

$$\mathcal{V}_{\mathcal{N}}(\gamma) := \frac{1}{G} \sum_{g=1}^G \text{tr} \left(\mathbb{E}_{\mathbb{P}(x_n)} [\tau(x_n) \tau(x_n)^\top] \text{Cov}_{\mathbb{P}(\lambda | \gamma, \mathbb{F})} (\eta_g(\gamma, \lambda)) \right).$$

If $N \mathbb{E}_{\mathbb{P}(\gamma | \mathbb{F})} [\bar{f}(\gamma) \mathcal{V}_{\mathcal{N}}(\gamma)]$ is strictly bounded away from 0 as $N \rightarrow \infty$, then

$$\sup_{\tilde{t} \in [0,1]} |\mathcal{E}(\tilde{t})| \rightarrow \infty \quad \text{in the Bayesian von-Mises expansion.}$$

More experimental results for Gamma–Poisson mixtures

We ran simulations of the Gamma–Poisson mixture with different ratios of N/G (average observations per group).

- When N/G is small:
 - IJ is biased significantly downwards
 - Bootstrap is biased somewhat downwards
- When N/G is larger:
 - Both improve
 - Both remain somewhat biased
 - The IJ and bootstrap perform similarly

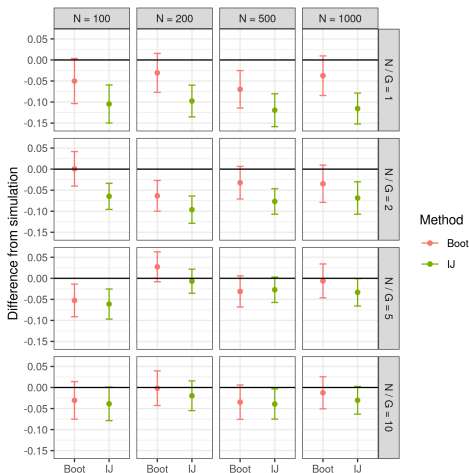


Figure 2: The error of the IJ and bootstrap covariances for different values of N and G . The y-axis shows the difference between $N(V - \hat{V}_{\text{sim}})$, where V is either \hat{V}_{IJ} or \hat{V}_{Boot} .

Exchangeable units. (A contradiction?)

Negative binomial observations.

Asymptotically linear in w .

Poisson observations with random effects.

Asymptotically non-linear in w .

Exchangeable units. (A contradiction?)

Negative binomial observations.

Poisson observations with random effects.

Asymptotically linear in w .

Asymptotically non-linear in w .

With a constant regressor, Gamma REs, and one RE per observation,
these are the same model, with the same $p(\gamma|X)$.

Is $\mathbb{E}_{p(\gamma|X,w)} [\gamma]$ linear in the data weights or not?

Exchangeable units. (A contradiction?)

Negative binomial observations.

Poisson observations with random effects.

Asymptotically linear in w .

Asymptotically non-linear in w .

With a constant regressor, Gamma REs, and one RE per observation, these are the same model, with the same $p(\gamma|X)$.

Is $\mathbb{E}_{p(\gamma|X,w)} [\gamma]$ linear in the **data weights** or not?

Exchangeable units. (A contradiction?)

Negative binomial observations.

Poisson observations with random effects.

Asymptotically linear in w .

Asymptotically non-linear in w .

With a constant regressor, Gamma REs, and one RE per observation, these are the same model, with the same $p(\gamma|X)$.

Is $\mathbb{E}_{p(\gamma|X,w)}[\gamma]$ linear in the **data weights** or not?

Trick question! We weight a log likelihood contribution, not a datapoint.

$$\log p(X|\gamma, w^m) = \sum_{n=1}^N w_n^m \log p(x_n|\gamma) \quad \log p(X|\gamma, \lambda, w^c) = \sum_{n=1}^N w_n^c \log p(x_n|\lambda, \gamma)$$

The two weightings are not equivalent in general.

What is the right exchangeable unit for a particular problem?

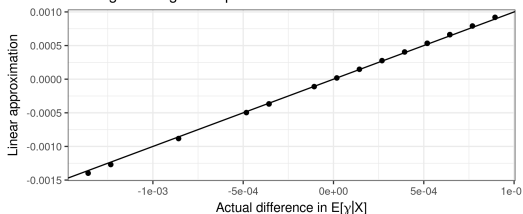
Exchangeable units: Experimental results revisited

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Uses $\log p(x_n|\gamma)$:

$$\psi_n = \mathbb{E}_{p(\gamma|X)} [\bar{\gamma} \bar{\ell}_n(\gamma)]$$

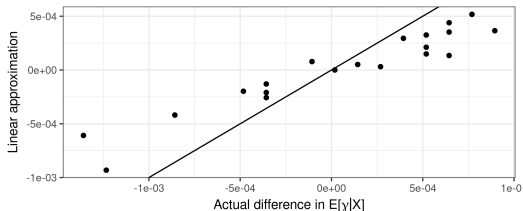
Negative Binomial model
leaving out single datapoints with $N = 800$



Uses $\log p(x_n|\gamma, \lambda)$:

$$\psi_n = \mathbb{E}_{p(\gamma, \lambda|X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)]$$

Poisson random effect model
leaving out single datapoints with $N = 800$



Exchangeable units: Experimental results revisited

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Uses $\log p(x_n | \gamma)$:

$$\psi_n = \mathbb{E}_{p(\gamma|X)} [\bar{\gamma} \bar{\ell}_n(\gamma)]$$

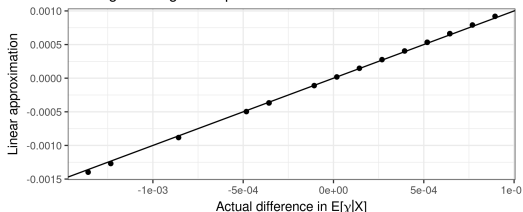
Not easily computable from
 $\gamma, \lambda \sim p(\gamma, \lambda | X)$
in general.

Uses $\log p(x_n | \gamma, \lambda)$:

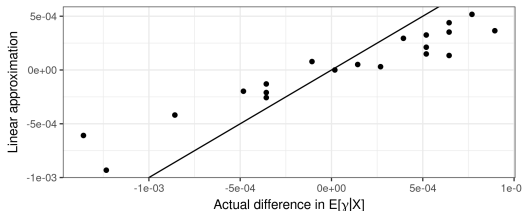
$$\psi_n = \mathbb{E}_{p(\gamma, \lambda | X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)]$$

Easily computable from
 $\gamma, \lambda \sim p(\gamma, \lambda | X)$.

Negative Binomial model
leaving out single datapoints with $N = 800$



Poisson random effect model
leaving out single datapoints with $N = 800$



Exchangeable units: Experimental results revisited

Our results were actually computed on **identical datasets** with $G = N$ and $g_n = n$.

Uses $\log p(x_n | \gamma)$:

$$\psi_n = \mathbb{E}_{p(\gamma|X)} [\bar{\gamma} \bar{\ell}_n(\gamma)]$$

Not easily computable from
 $\gamma, \lambda \sim p(\gamma, \lambda|X)$
in general.

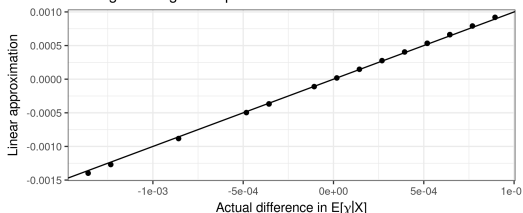
Uses $\log p(x_n | \gamma, \lambda)$:

$$\psi_n = \mathbb{E}_{p(\gamma, \lambda|X)} [\bar{\gamma} \bar{\ell}_n(\gamma, \lambda)]$$

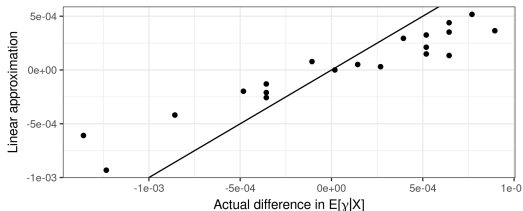
Easily computable from
 $\gamma, \lambda \sim p(\gamma, \lambda|X)$.

May still be useful when $p(\lambda|X)$
is *somewhat* concentrated.

Negative Binomial model
leaving out single datapoints with $N = 800$



Poisson random effect model
leaving out single datapoints with $N = 800$



Observations and consequences

- For finite-dimensional models which concentrate asymptotically:
 - Posterior expectations are approximately linear in data weights
 - The linearized variance estimate (infinitesimal jackknife) is consistent
 - The residual of the von Mises expansion vanishes
- For high-dimensional models which marginally concentrate only asymptotically:
 - Posterior expectations are not approximately linear in data weights
 - The linearized variance estimate (infinitesimal jackknife) is inconsistent
 - The residual of the von Mises expansion does not vanish
 - Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice.
... Especially given the linear approximation's huge computational advantage.

Observations and consequences

- For finite-dimensional models which concentrate asymptotically:
 - Posterior expectations are approximately linear in data weights
 - The linearized variance estimate (infinitesimal jackknife) is consistent
 - The residual of the von Mises expansion vanishes
- For high-dimensional models which marginally concentrate only asymptotically:
 - Posterior expectations are not approximately linear in data weights
 - The linearized variance estimate (infinitesimal jackknife) is inconsistent
 - The residual of the von Mises expansion does not vanish
 - Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice.
... Especially given the linear approximation's huge computational advantage.
- When the weighting is linear, there are many other applications:
 - Cross-validation
 - Conformal inference
 - Identification of influential subsets
- When the weighting is non-linear, the inconsistency results should apply more widely:
 - The EM algorithm
 - The nonparametric bootstrap
 - Local prior sensitivity measures

Observations and consequences

- For finite-dimensional models which concentrate asymptotically:
 - Posterior expectations are approximately linear in data weights
 - The linearized variance estimate (infinitesimal jackknife) is consistent
 - The residual of the von Mises expansion vanishes
- For high-dimensional models which marginally concentrate only asymptotically:
 - Posterior expectations are not approximately linear in data weights
 - The linearized variance estimate (infinitesimal jackknife) is inconsistent
 - The residual of the von Mises expansion does not vanish
 - Even if the error $\mathcal{E}(w)$ does not vanish, it can still be small enough in practice.
... Especially given the linear approximation's huge computational advantage.
- When the weighting is linear, there are many other applications:
 - Cross-validation
 - Conformal inference
 - Identification of influential subsets
- When the weighting is non-linear, the inconsistency results should apply more widely:
 - The EM algorithm
 - The nonparametric bootstrap
 - Local prior sensitivity measures

Preprint: Giordano and Broderick [2023] ([arXiv:2305.06466](https://arxiv.org/abs/2305.06466))

(Major update in progress, coming soon.)

- A. Gelman and M. Heidemanns. The Economist: Forecasting the US elections., 2020. URL <https://projects.economist.com/us-2020-forecast/president>. Data and model accessed Oct., 2020.
- A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.
- R. Giordano and T. Broderick. The Bayesian infinitesimal jackknife for variance. *arXiv preprint arXiv:2305.06466*, 2023.
- J. Huggins and J. Miller. Reproducible model selection using bagged posteriors. *Bayesian Analysis*, 18(1):79–104, 2023.
- R. Kass, L. Tierney, and J. Kadane. The validity of posterior expansions based on Laplace’s method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, 1990.
- B. Kleijn and A. Van der Vaart. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6: 354–381, 2012.